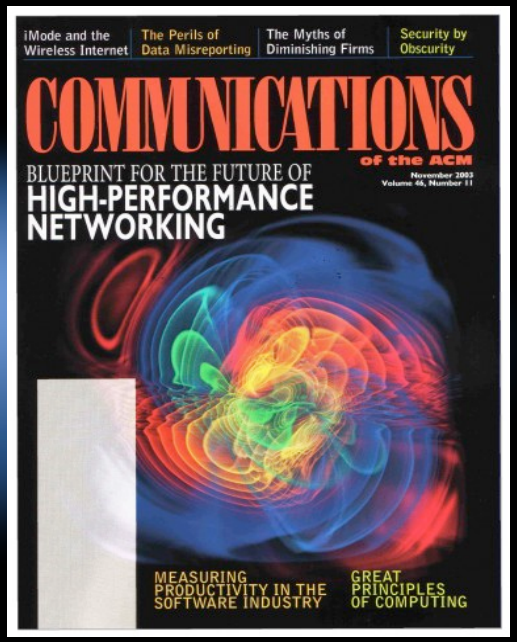
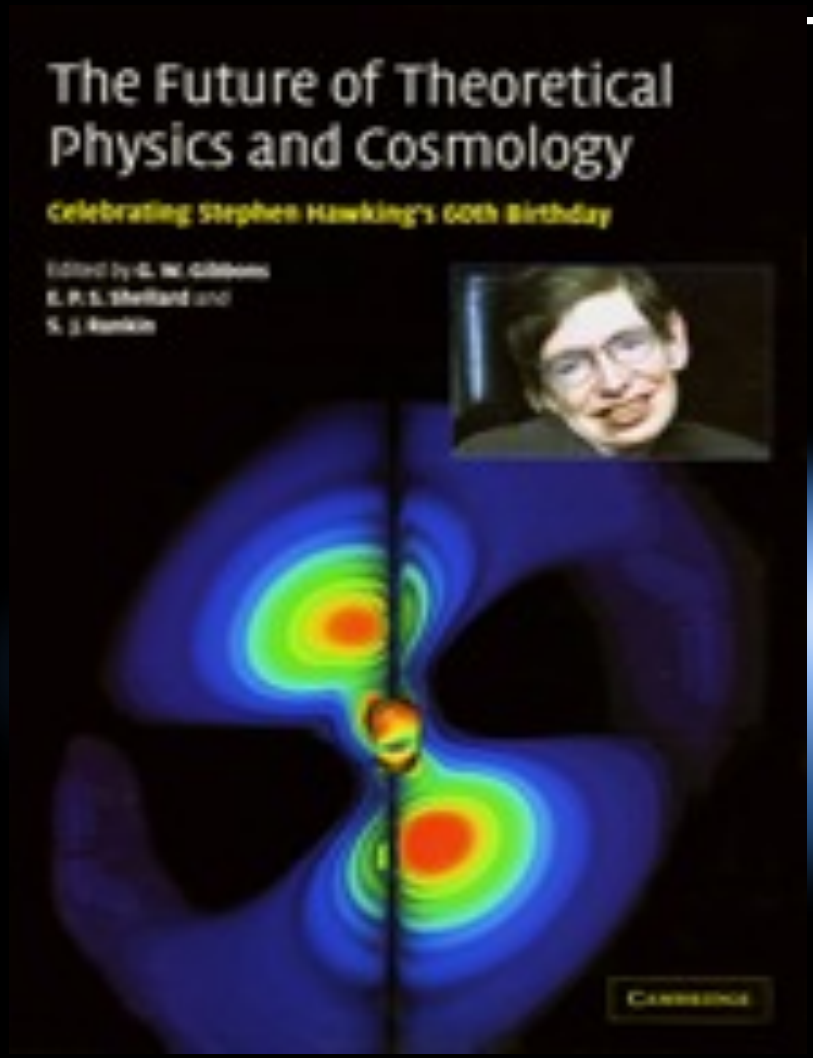
Transformation of Science through Cyberinfrastructure

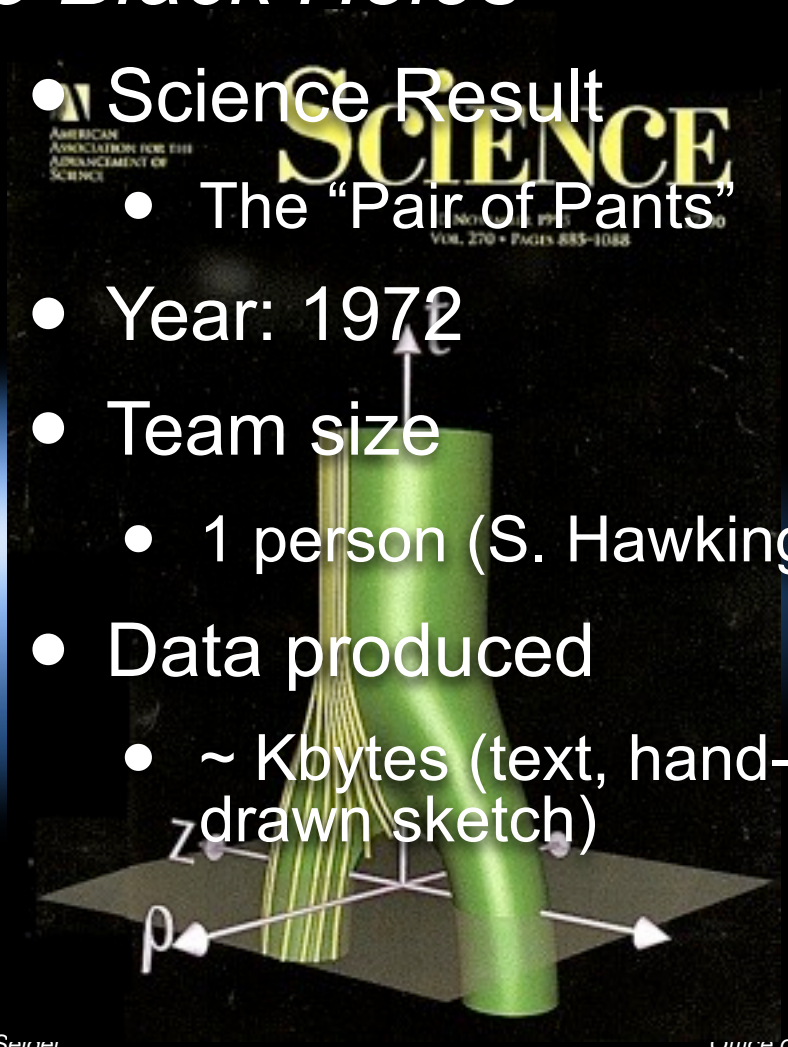
Edward Seidel
Director, Office of Cyberinfrastructure
National Science Foundation
hseidel@nsf.gov



Transformation of Science

Two Black Holes



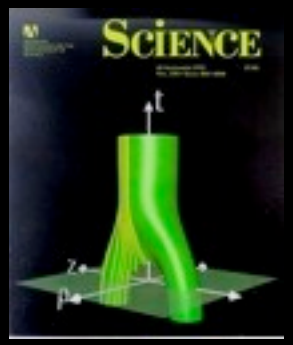


- Science Result
  - The "Pair of Pants"
- Year: 1972
- Team size
  - 1 person (S. Hawking)
- Data produced
  - ~ Kbytes (text, hand-drawn sketch)

Data-Driven Multiscale Collaborations* for Complexity

Great Challenges of 21st Century

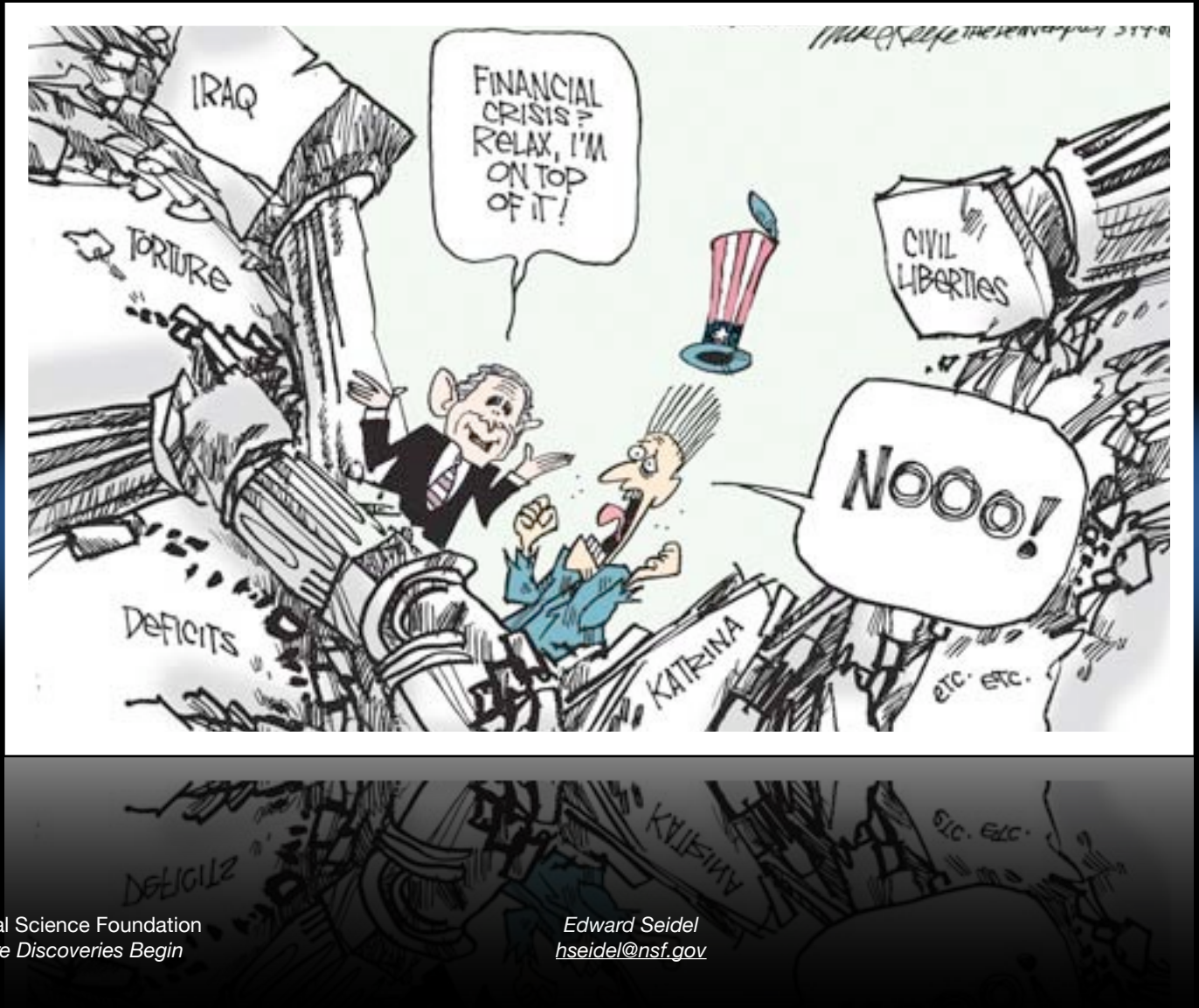
- Multiscale Collaborations
  - General Relativity, Particles, Geosciences, Bio, Social...
  - And all combinations...
- Science and Society being transformed by CI and Data
  - Completely new methodologies
  - "The End of Science" (as we know it)
- CI plays central role
  - No community can attack challenges
  - Technical, CS, social issues to solve
- *Places requirements on computing, software, networks, tools, etc*



*Small groups still important!

Edward Seidel
hseidel@nsf.gov

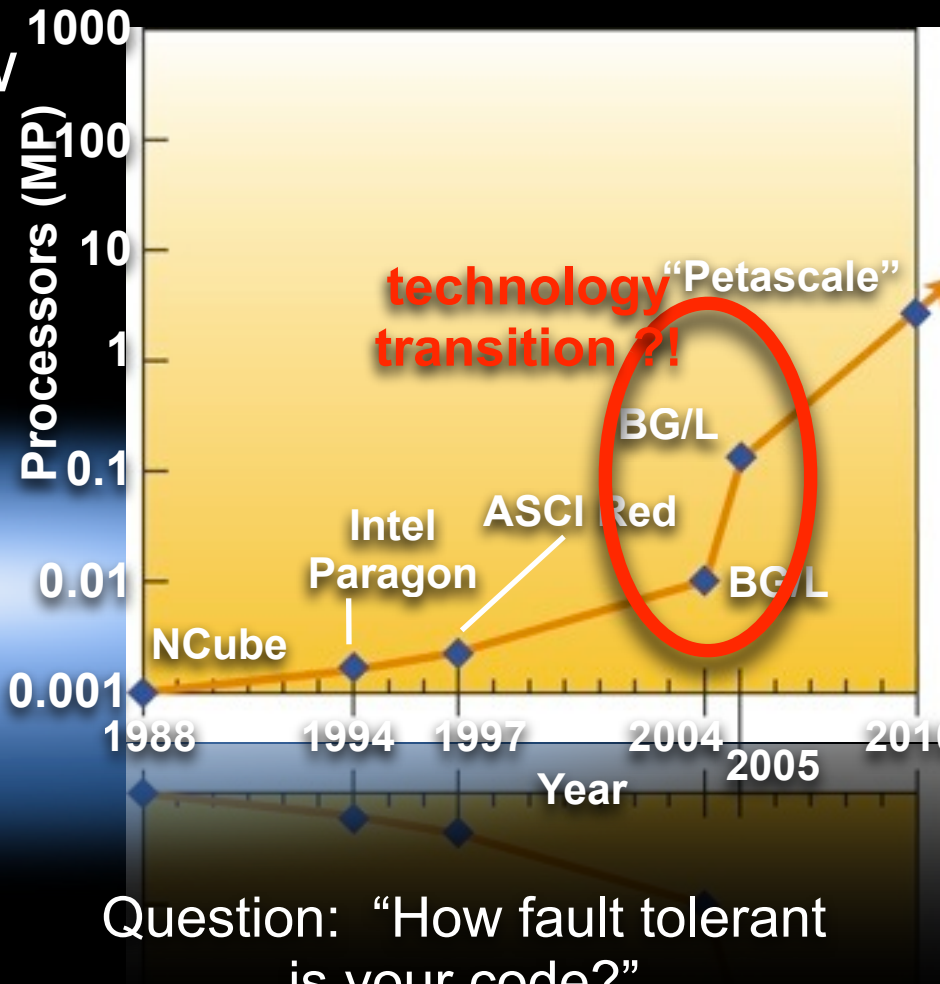
Crises to Deal With



Edward Seidel
hseidel@nsf.gov

Technology Crisis
(Adapted from Simon/Shalf)

- "The processor is the new transistor" (Patterson)
  - BG/L at LLNL: as many procs as transistors in the MC68000
  - N_{procs} has made a transition
  - Programming parallel codes like assembly language moving bits between transistors
- Multicore is coming on fast
  - N_{cores} doubles every 18 months, clock rate more or less fixed
  - Programming: MPI within a chip???, No: MPI + ???
- Many CS, CI Challenges



Question: "How fault tolerant is your code?"

Edward Seidel
hseidel@nsf.gov

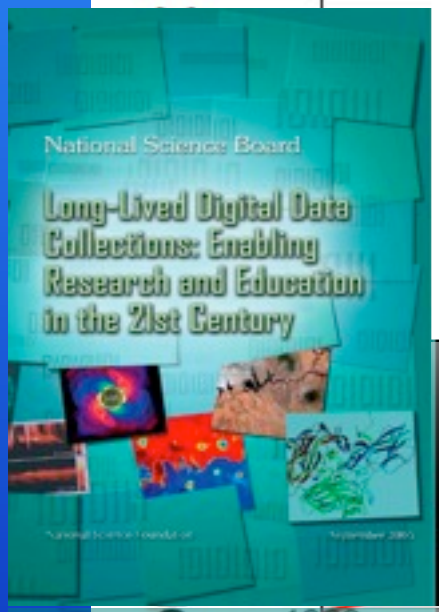
Data Crisis: Information Big Bang

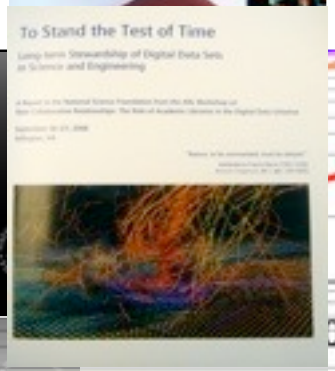
PCAST Digital Data

Industry

Wired, Nature

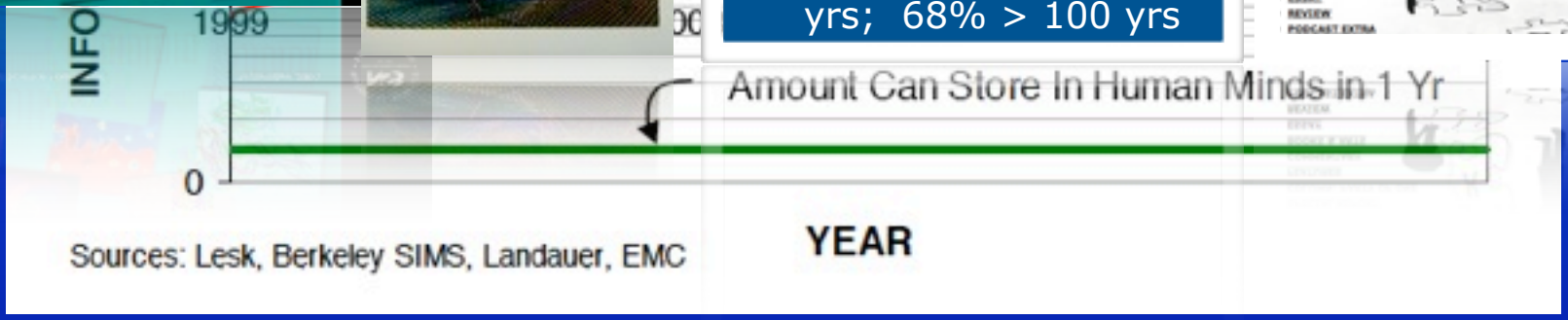
NSB Report: Long-Lived Digital Data Collections Enabling Research and Education in the 21st Centur



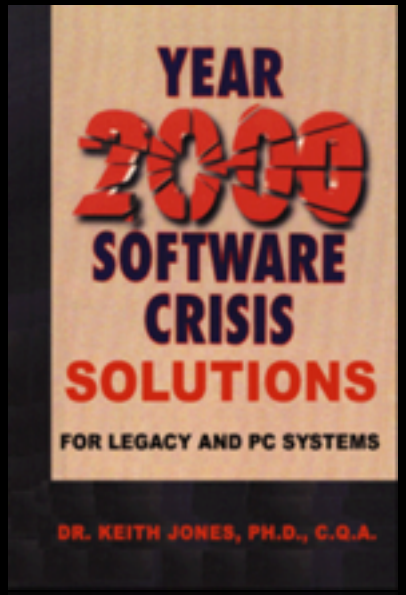


Storage Networking Industry Association (SNIA) *100 Year Archive Requirements Survey Report*

"there is a pending crisis in archiving… we have to create long-term methods for preserving information, for making it available for analysis in the future." 80% respondents: >50 yrs; 68% > 100 yrs

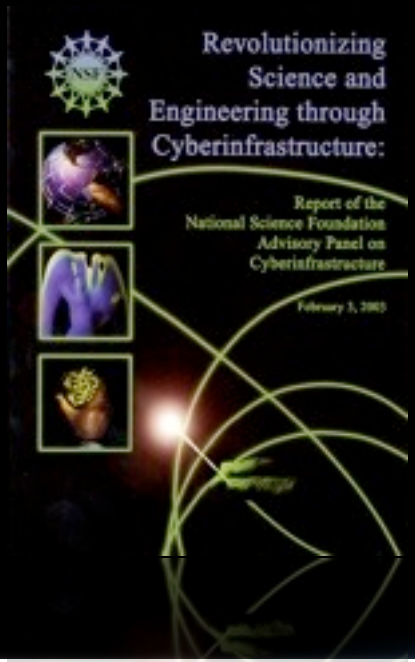


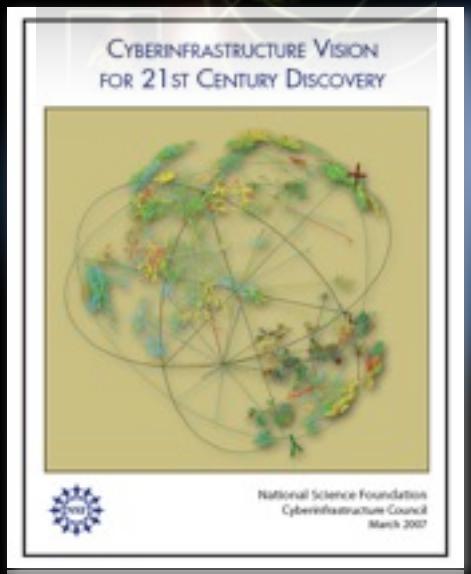
Software Crisis



- Computers are exceedingly complex
  - Desktops with hundreds of cores
  - Supercomputers with millions of cores
  - They last 3-4 years...
- Software systems and applications
  - Science apps have 10^3 to 10^{6+} lines, have bugs
  - Applications may take decades to develop
  - We spend at least 10x as much on hardware
  - *GC communities place requirements on software for complex CI (not just HPC!)*
- We have a *crisis* in software
  - We don't know how to write it!
  - Is our science reproducible? If not...not science!

Toolkit for complex CI?





NSF Vision

"National-level, integrated system of hardware, software, data resources & services... to enable new paradigms of science"

Virtual Organizations for Distributed Communities

High Performance Computing

Data & Visualization/ Interaction

Learning & Work Force Needs & Opportunities

Edward Seidel
hseidel@nsf.gov

Office of Cyberinfrastructure

National Science Foundation
Where Discoveries Begin

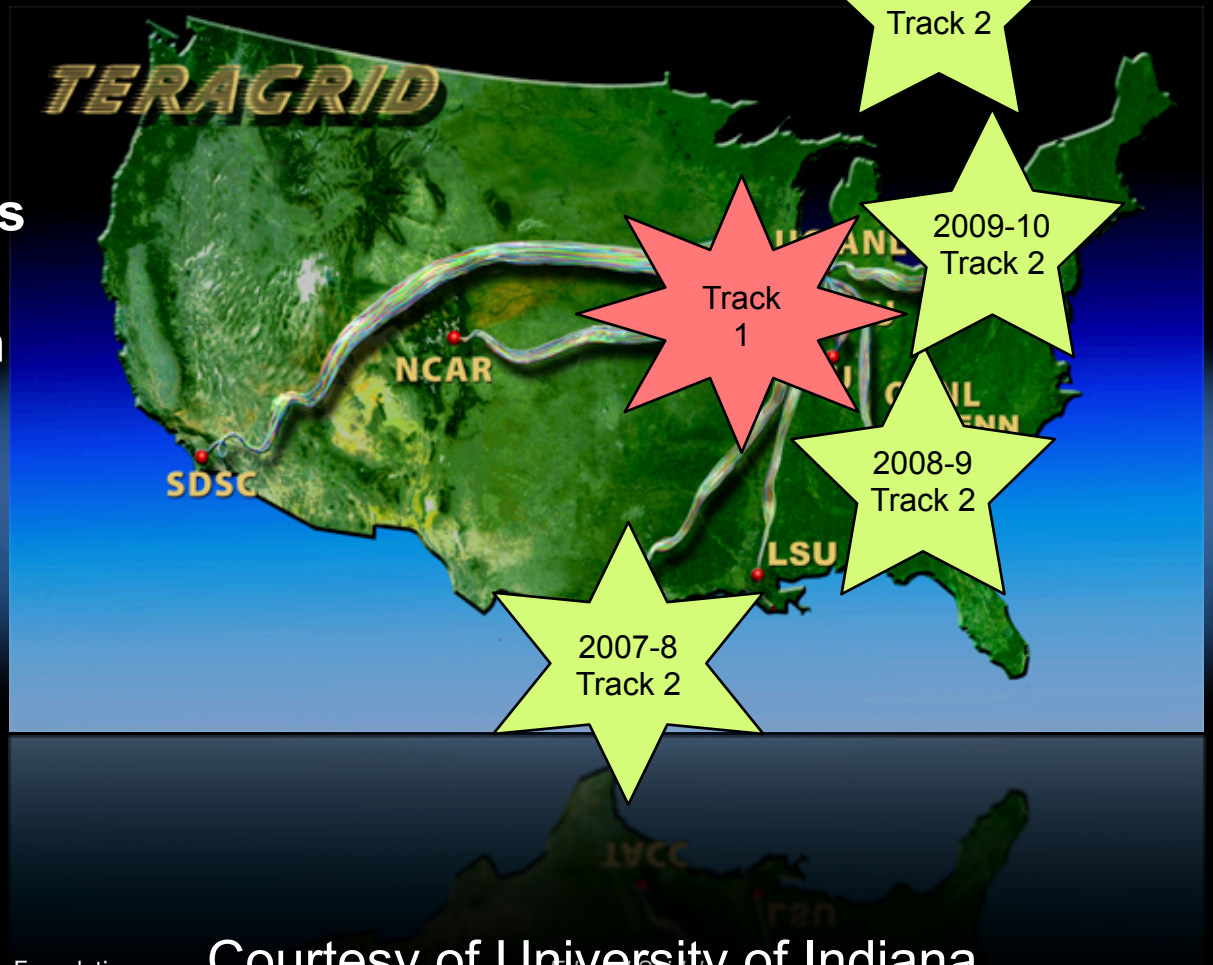
Shared Resource Environments

Computers

Data services

Visualization services

People



Modeling and simulation

Data analysis & visualization

User support

Training

Common user environments

Tools for educators

Science Gateways

Courtesy of University of Indiana

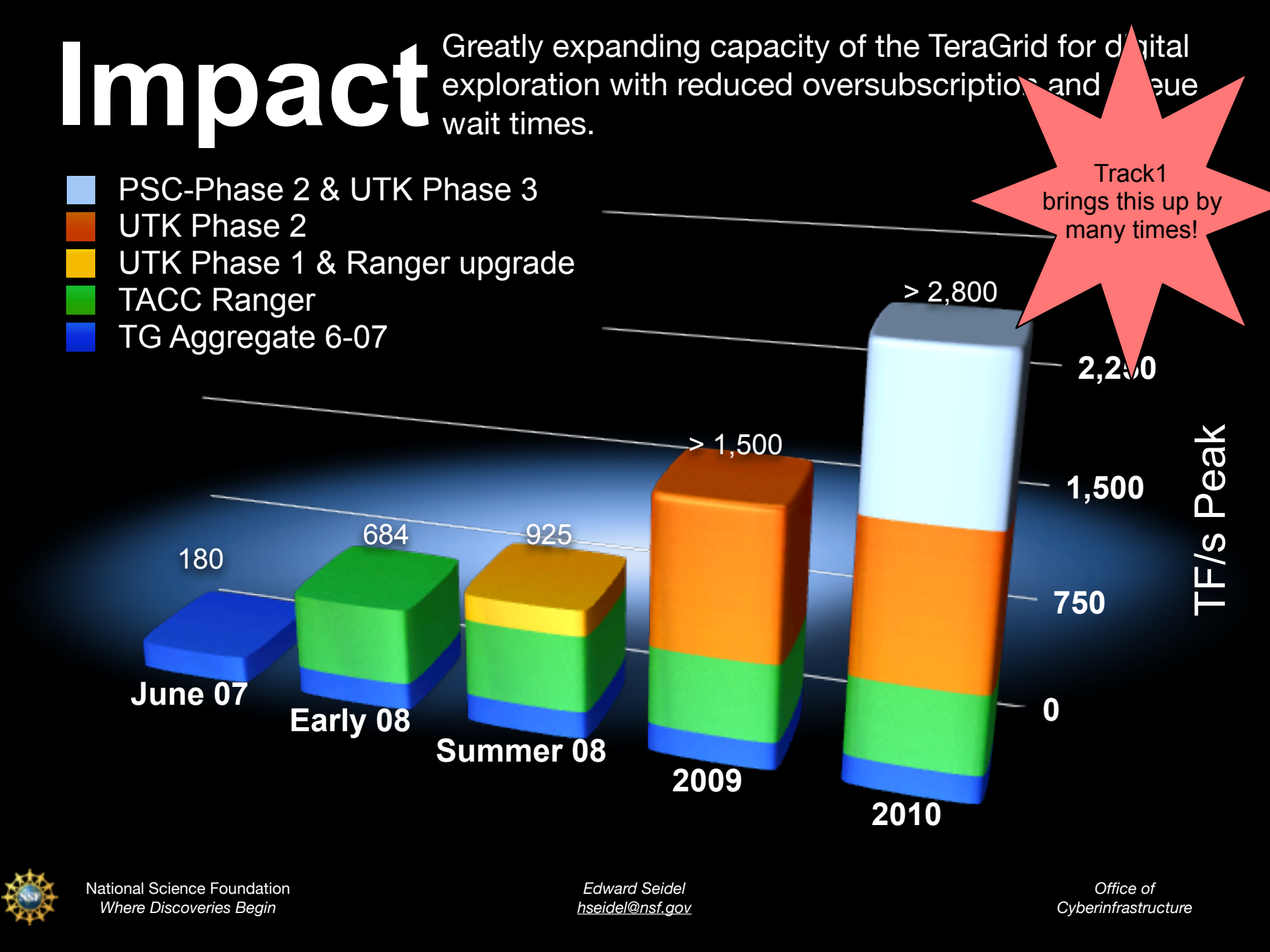
National Science Foundation
Where Discoveries Begin

Edward Seidel
hseidel@nsf.gov

Office of Cyberinfrastructure

Impact

Greatly expanding capacity of the TeraGrid for digital exploration with reduced oversubscription and queue wait times.



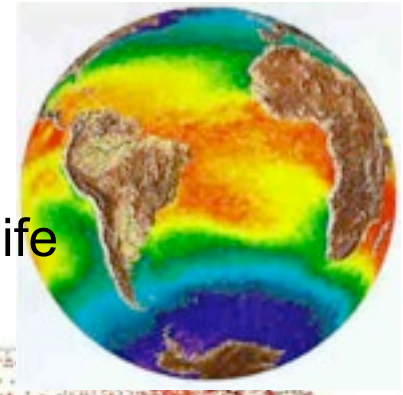
Edward Seidel
hseidel@nsf.gov
Office of Cyberinfrastructure

Track 1

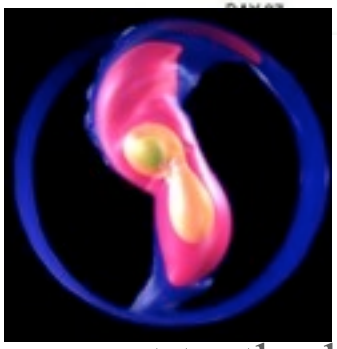
Blue Waters Project

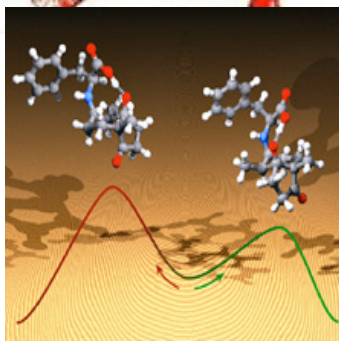
Blue Waters Petascale System (2011!)

- **Blue Waters General Characteristics**
  - Based on IBM PERCS
  - 1 petaflops *sustained* performance on real apps
- **Blue Waters System Characteristics**
  - > 200,000 cores; multicore POWER7 processors
  - > 32 gigabytes of main memory per SMP
  - > 10 petabytes of user disk storage
  - > 100 Gbps external connectivity (initial)
  - Fortran, Co-Array Fortran, C/C++, UPC, MPI/MPI2, OpenMP, Cactus, Charm++
- **Blue Waters Interim Systems at NCSA**
  - POWER 5+/6 software and application development testbeds
- **Blue Waters System Training and Support**



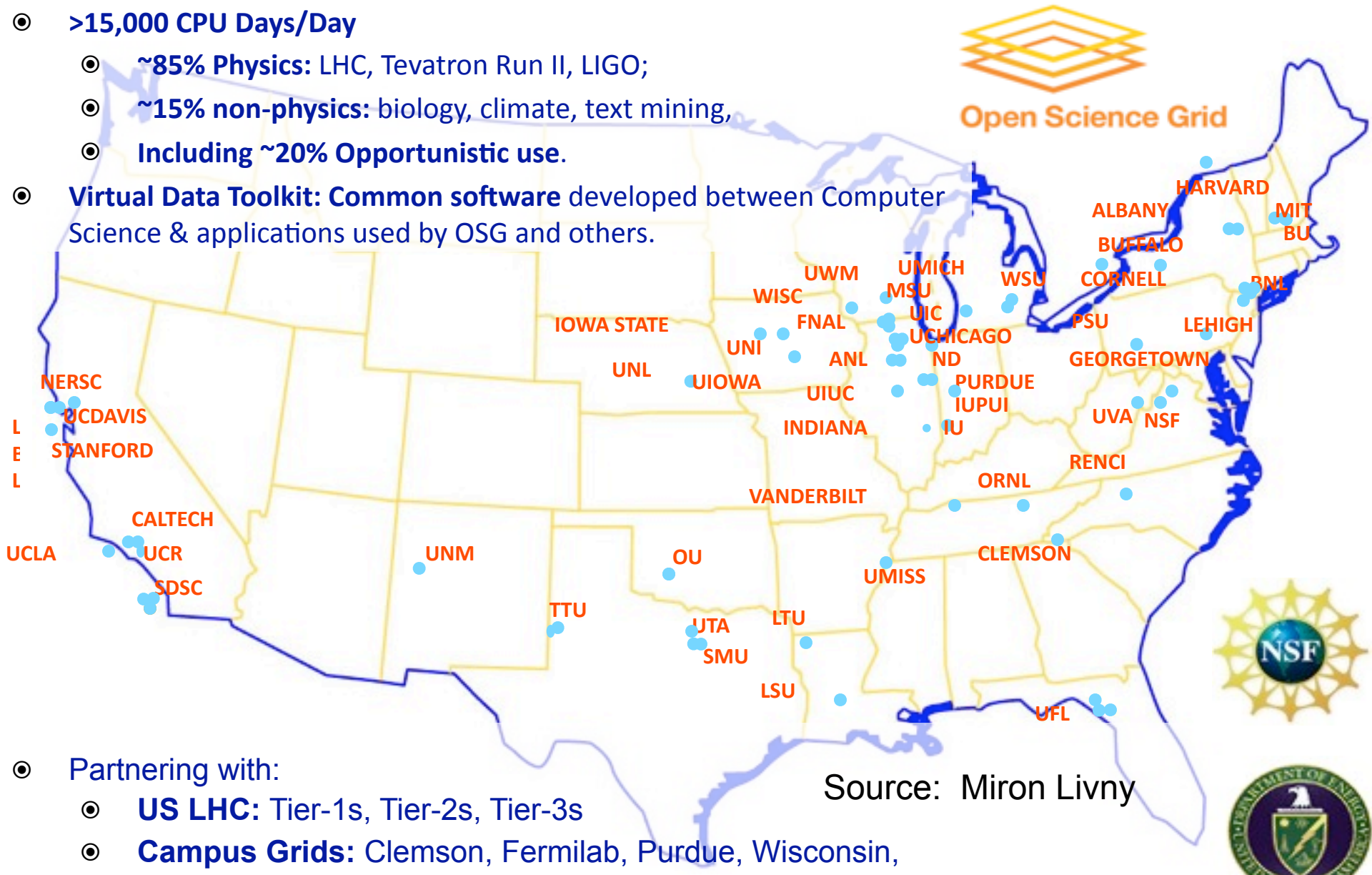
Earth & Life Sciences





Astronomy and Molecular Science

- **Access to 45,000 Cores, 6 Petabytes Disk, 15 Petabytes Tape**
- **>15,000 CPU Days/Day**
  - **~85% Physics:** LHC, Tevatron Run II, LIGO;
  - **~15% non-physics:** biology, climate, text mining,
  - **Including ~20% Opportunistic use**.
- **Virtual Data Toolkit: Common software** developed between Computer Science & applications used by OSG and others.



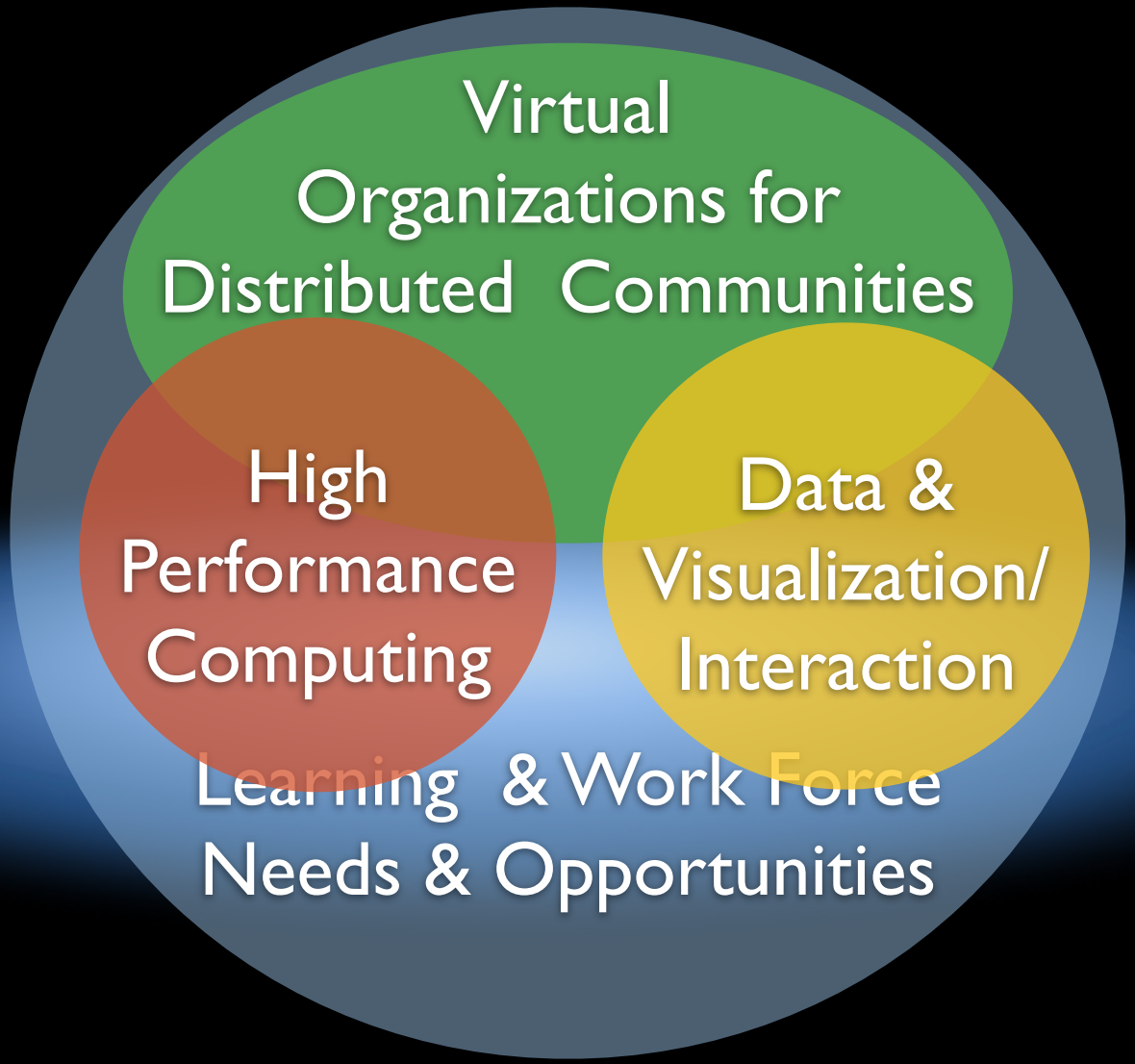
Source: Miron Livny

- Partnering with:
  - **US LHC:** Tier-1s, Tier-2s, Tier-3s
  - **Campus Grids:** Clemson, Fermilab, Purdue, Wisconsin,
  - **Regional & National Grids:** TeraGrid, New York State Grid, EGEE, UK NGS
  - **International Collaboration:** South America, Central America, Taiwan, Korea, UK.

Open Science Grid as Model "Campus Bridge"

- NSF very interested in creating "bridges" from campus to national CI
- OSG is a *national* CI, *locally* deployed…good model
- We are very interested in…
  - Exploring ways to integrate campuses better with national centers, instruments
  - TeraGrid-OSG cooperation
    - Driven by applications!
  - Understanding example science communities that can benefit from, drive this: GC Communities will require
  - Related international cooperation: EGEE/EGI, etc

Edward Seidel
hseidel@nsf.gov



Edward Seidel
hseidel@nsf.gov

Data, Data Analysis, and Visualization

- "Any cogent plan must add
  phenomenal growth of data
- Goals are to
  - Catalyze the development of a
    and engineering data collecti
    extensible, and evolvable
  - Support development of a new
    tools and services for data discovery, integration,
    visualization, analysis and preservation
- The resulting national digital data framework
  will be an integral component in national C

Where are the data in this paper? I can't repeat their results!

I can't repeat my results!

I can't visualize my results!

DataNet: $100M 5 year program

National CI Blueprint

Clouds?

Exascale (multi-agency?)

Nets

Software

Algorithms

Track 1

DataNet

I need all of this to start to solve my problem!

Apps

Track 2

Track 2

Grand Challenge Community

DataNet

VO

DataNet

Apps

LC

Apps

Campus

Campus

Campus

Campus

Nets (campus)

Computational Science
PITAC Report Summary

“Itself a discipline, computational science serves to advance all of science … inadequate and outmoded structures within the Federal government and the academy today do not effectively support this critical multidisciplinary field”

“Universities must significantly change organizational structures: multidisciplinary & collaborative research are needed [for US] to remain competitive in global science”

“Together with theory

REPORT TO THE PRESIDENT
JUNE 2005
COMPUTATIONAL SCIENCE: ENSURING AMERICA'S COMPETITIVENESS
PRESIDENT'S INFORMATION TECHNOLOGY ADVISORY COMMITTEE

Task Force OCI Contacts

Campus Bridging

Jennifer Schopf
jschopf@nsf.gov

Data & Viz

Jon Stoffel
jstoffel@nsf.gov

Software

Abani Patra
apatra@nsf.gov

HPC (Clouds Grids)

Rob Pennington
rpenning@nsf.gov

Barry Schneider
bschneid@nsf.gov

Education Workforce

Rob Pennington
rpenning@nsf.gov

Grand Challenge VOs